基于 PageRank 的多维度微博用户影响力度量 *

罗　芳[1]，徐　阳[1]，蒲秋梅[2]，邱奇志[1]
(1. 武汉理工大学 计算机科学与技术学院，武汉 430063; 2. 中央民族大学 信息工程学院，北京 100081)

摘　要：近年来社交网络的发展推动了多个领域的研究，如舆情监控、广告推荐、意见领袖识别等，而社交网络用户的影响力度量则是以上研究的基础。以新浪微博为研究对象，旨在提出一种适用性更广、考虑因素更全面的微博用户影响力度量算法，将用户基本属性、用户交互行为和用户博文内容三个维度因素融入传统 PageRank 算法中，提出了一种多维度微博用户影响力度量算法——MDIR(multi-dimension influence rank)。实验结果表明，MDIR 算法相较于其他常用的五种影响力度量算法，能更加全面、真实地反映微博用户的实际影响力。
关键词：微博；用户影响力；PageRank；用户行为
中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.10.0798

Multi-dimensional measure of Microblog user influence based on PageRank

Luo Fang[1], Xu Yang[1], Pu Qiumei[2], Qiu Qizhi[1]
(1. *School of Computer Science & Technology*, *Wuhan University of Technology*, *WuHan* 430063, *China*; 2. *School of Information Engineering*, *Minzu University of China*, *Beijing* 100081, *China*)

Abstract: In recent years, the development of social networks had promoted research in many fields, such as public opinion monitoring, advertising recommendation and opinion leader identification etc. The influence measurement of social network users is the basis of the above research. This paper integrated the basic attributes of user, interaction behavior of user and user's microblog content into the PageRank algorithm, therefore, it proposed a multi-dimensional user influence measurement algorithm:MDIR(multi-dimension influence rank). The experiment shows that, the MDIR can reflect the actual influence of microblog users more comprehensively and realistically than other five commonly used influence measurement algorithms.
Key words: Microblog; user influence; PageRank; user behavior

0　引言

随着互联网技术的迅速发展，以博客技术为代表，围绕用户互动与个性体验的互联网应用技术进一步推动了以开放、共享为特征的 Web 2.0 时代向具有信息融合特征的 Web 3.0 时代过渡。微博是微型博客的简称，是一种基于关注机制分享简短实时 信息的广播式的社交网络平台。据 CNNIC 发布的第 41 次《中国互联网络发展状况统计报告》中的数据显示[1]，截至 2017 年 12 月，我国微博用户规模为 3.76 亿，推动用户使用率持续增长达到 40.9%，较 2016 年 12 月上升 3.8 个百分点。微博平台中每个用户除了发布自己原创的微博信息外，还可以随意的转发、评论、点赞其他用户的微博信息，不同用户之间的相互转发、评论、点赞等行为促成了微博信息传播网络的形成。另外，微博平台还具有用户使用门槛较低、微博内容短小精悍、可阅读性强的特点，这些特点使得微博信息的传播速度更快以及影响范围更广。

微博用户的影响力可以理解成是某用户发布微博后引起其他用户行为改变的能力。在微博信息传播的过程中，不同影响力的用户对微博信息的操作（如转发、评论等）和态度（如支持、反对等），会对微博信息的传播范围与传播深度产生不同的影响。用户影响力的度量在网络舆情监控、广告投放、用户推荐等领域有着重要的应用。对于舆情监控而言，监控过程中需要对某些高影响力的用户采取特别措施，以免舆情泛滥；对于广告投放而言，选取高影响力的用户作为初始的广告传播的中心可以使得传播效果最大化；对于用户推荐而言，用户感兴趣领域的“意见领袖”[2]常常都是默认推荐的对象。综上所述，用户影响力的度量在当前的热点研究中扮演着不可忽视的角色。

为了合理地度量微博用户的影响力，本文在 PageRank 算法的基础上进行改进，并提出了 MDIR（multi-dimension influence rank）算法，相比较于其他影响力度量算法而言，MDIR 算法考虑的影响因素更为全面、合理，其得到的用户影响力排名也更为客观。

1　相关工作

目前，国内外学者对于社交网络用户影响力度量的研究一般基于用户的基本属性、交互行为以及发布的博文内容三个方面：

a) 基于用户基本属性的影响力度量方法。

用户的基本属性是用户影响力最原始的体现，常见的基本属性如粉丝数、发博数等，这些大多都是当前流行的影响力度量算法所考虑的特征因素。Cha 等人[3]选取 Twitter 中用户的粉丝数、转发数、评论次数三个属性，按节点度计算用户的影响力并对比所得结果的相关性，其实验结果表明用户

收稿日期：2018-10-24；**修回日期：**2018-12-27　**基金项目：**国家教育部人文社会科学研究规划基金资助项目（18YJAZH087）
作者简介：罗芳（1977-），女，湖北天门人，副教授，硕导，博士，主要研究方向为数据挖掘、自然语言处理（luof@whut.edu.cn）；徐阳（1994-），男，硕士，主要研究方向为社会计算、数据挖掘；蒲秋梅（1976-），女，讲师，硕导，博士，主要研究方向为智能系统、社会计算；邱奇志（1969-），女，副教授，硕导，博士，主要研究方向为智能计算、自然语言处理.

粉丝数的多少与用户微博被转发数、被评论数并不成正比例关系。Mao 等人[4]对微博用户的活跃度进行了分析，但是其仅仅只通过评论数进行分析，并没有分析其他因素，也没有排除微博平台中“僵尸粉”对用户影响力的干扰。

b) 基于用户交互行为的影响力度量方法。

用户间的交互行为是用户影响力的直接体现，常见的用户交互行为有转发、评论、提及、点赞等。张昊等人[5]基于用户的基本属性与交互行为提出了 UI 算法，其首先引出了“用户影响力”与“用户被影响力”的概念，“用户影响力”是基于用户的粉丝数、发博数、微博被转发数等属性计算得来，而“用户被影响力”则是基于粉丝用户与关注用户之间的交互情况，如粉丝用户对某关注用户的微博进行转发、评论的次数占该粉丝用户对其所有关注用户的微博进行转发、评论的总数的百分比，但是作者对用户各种交互行为采取的量化手段是归一化，这并不符合实际微博传播情况；王顶等人[6]在张昊等人研究的基础上考虑了不同行为的权重值，其实验部分也从数据出发对排序结果作出了详细的论证，但是其只基于用户间的关注关系构建网络拓扑，没有考虑到微博中存在大量的“僵尸粉”，即关注了某用户之后并不对其产生交互行为，这类粉丝在微博的传播中起不到作用；孙红等人[7]综合考虑用户的实际微博活动行为以及微博网络的拓扑结构，进而提出了 MBUI-Rank 算法，其实验结果表明该算法计算出的用户影响力较为准确与客观；齐超等人[8]在 PageRank 算法的基础上加以改进提出了 BWPR 算法，分别对转发、评论、提及三种行为构建拓扑网络，虽然该算法在实验中得到了很好的效果，但是其只考虑了用户交互行为这种显性特征，没有考虑隐性的用户兴趣偏好。

c) 基于博文内容的影响力度量方法。

用户所发布的微博携带着大量信息，通过对博文内容进行分析可以获取用户所感兴趣的话题或者用户的情感属性，这些特征也被广泛地应用在影响力度量的领域中。Weng 等人[9]基于 Twitter 数据提出了 TwitterRank 算法，其不仅考虑网络结构，而且基于推文内容分析了每个用户所发推文的话题相似性，最后将用户在每个主题中的影响力值求和得到用户在整个网络中的影响力值。虽然此方法在实验中得到了不错的效果，但是其也只考虑了推文数目和话题相似度，忽略了用户之间的交互行为特征。师亚凯等人[10]通过引入网络拓扑结构中用户行为与基于博文内容的用户兴趣相似度来衡量用户间的影响力，但是没有考虑用户的粉丝数、认证情况等用户基本属性。

鉴于当前研究的不足之处，本文在 PageRank 算法的基础上进行改进，提出了一种融合用户基本属性、用户交互行为、用户博文内容的多维度微博用户影响力度量算法——MDIR（multi-dimensionin fluence rank）。

2 MDIR 算法基本原理

2.1 PageRank 算法

PageRank 算法是互联网中广泛用于网页排名的经典算法[11]，其核心思想是研究网络的拓扑结构并计算页面中的入链数（即页面链接的次数），从而确定该页面的排名顺序。PageRank 算法的相关公式如式（1）所示。其中：$PR(p_i)$ 为页面 p_i 的 PageRank 值；$I(p_i)$ 为页面 p_i 的入链集合；$|O(p_j)|$ 为页面 p_j 的出链集合中网页的数量；d 是阻尼系数，通常采用 d = 0.85[12]。

$$PR(p_i)=(1-d)+d\sum_{p_j\in I(p_i)}\frac{PR(p_j)}{|O(p_j)|} \tag{1}$$

传统的 PageRank 算法用于微博用户的影响力度量时，将微博平台中的用户类比为 Web 网络中的网页，仅仅考虑了用户与用户之间的关注与被关注关系。若直接采用 PageRank 算法度量用户影响力会存在以下问题：

a) 初始 PR 值的确定不够客观。

PageRank 算法对于网页的初始 PR 值采取的计算方法是取平均数，这样的方法并不适用于微博用户影响力度量。由于每个用户间的差异（具体表现在用户的粉丝数、发博数、是否认证等），若直接采用取平均数的方式得到用户的初始 PR 值，则会忽略用户自身属性对微博传播所带来的影响，从而导致最后计算出的用户影响力排名不够客观。

b) PR 值的分配方式不够合理。

网页将自身 PageRank 值均匀地分给它所链出的网页，这样的计算方式对于微博用户显然不合理。因为不是所有用户都能做到对他们关注的所有用户一视同仁，大多数用户只对其中的一部分表现出兴趣从而愿意为其投入更多的注意力。

c) 网络结构的考虑不够全面。

PageRank 算法在微博中的应用是基于用户的关注关系，由此算法很大程度上依赖于粉丝数。但由于粉丝中可能包含很多无效的僵尸粉与沉默粉，所以粉丝数并不能真实地反映用户的影响力。此外，微博用户是动态的，他们之间存在着各种各样的行为，如转发微博、评论微博等。这些行为对用户微博的传播有着明显的推动作用，而且这些行为不仅仅发生在有关注关系的用户之间，如新浪微博推出的“微吧”和“微话题”功能，用户不需关注即可获取微吧或微话题内其他用户所发的消息并且可以对其进行转发、评论等操作。

2.2 MDIR 影响力度量因素

针对上文分析的 PageRank 算法在度量用户影响力时存在的问题，本文提出的 MDIR 算法综合考虑了以下三个维度的因素：

a) 基于用户基本属性的影响因素。

用户自身的基本属性如用户的粉丝数、发博数、认证情况。用户的粉丝数、发博数对于用户的影响力而言是一个比较直观的影响因素。通常来讲，用户的粉丝数越多，意味着能看到该用户发布的微博信息的用户也越多，从而导致该用户所能影响到的人数也越多；如果说用户的粉丝数决定了该用户影响力的范围，那么用户的发博数则决定了该用户影响力的深度。对于相同规模的粉丝群体，用户的发博数越多，则每个粉丝被该用户微博信息所影响到的次数也会随之增多，即该用户影响力的深度越深；用户的认证情况对用户的影响力而言是一个潜在的影响因素，如果用户通过新浪微博认证，则其微博的可信度越高，从而使得该微博被评论和被转发的可能性越大，因此用户影响力越大。

b) 基于用户间交互行为的影响因素。

在微博平台中，用户所发布、转发的微博都以潜移默化的方式影响着他们的粉丝，而粉丝的评论、转发、提及等行为则又反过来推动微博信息的传播。微博信息被转发的次数越多，意味着该微博信息被传播得越广；微博信息被评论的次数越多，意味着该微博信息受关注度越高；用户被提及次数越多，说明该名用户在其粉丝群体中建立的威信度越高。

c） 基于用户博文内容的影响因素。

用户发布与转发的微博信息可以看做是用户个人兴趣的体现，通过对博文内容的挖掘可以得到每条微博的主题分布向量，而综合一个用户所有微博的主题分布向量，可以得到代表一个用户兴趣的主题分布向量，本文简称为兴趣分布向量。Weng 等人[9]指出，在 Twitter 中用户感兴趣的话题越相

似，则他们相互关注的可能性越大。虽然其在 Twitter 平台下做的实验，但由于 Twitter 与微博的结构具有相似性，其结果也可以被用于微博平台。李志宏等人[13]指出微博用户间具有“同好性”，即用户更倾向于与自己有相同兴趣爱好的用户建立关系（转发关系、评论关系等）。

综上所述，本文提出的 MDIR 算法基于用户的基本属性，计算出用户的初始影响力值，解决了传统 PageRank 算法的不够客观的缺点；基于用户间的交互行为与用户博文内容，计算用户间的传播意愿，从而改进了传统 PageRank 的影响力均值分配的问题；基于用户的交互行为，构建用户微博信息的传播网络，有效地排除了大量“僵尸粉”“沉默粉”对用户影响力的干扰。

3　MDIR 算法具体实现

3.1　微博传播网络的构建

根据 2.2 节所述，本文通过用户的转发、评论、提及行为构建一个微博信息的传播网络简称微博传播网络，其定义如下：

定义 1 微博传播网络。设微博传播网络为 $G=(V, E, B)$，其中 $V=\{v_i | i=1,2,3,\dots,n\}$ 为微博传播网络中节点的集合，v_i 代表有涉及到转发、评论、提及三种行为中一种或多种的微博用户；$E=\{(v_i,v_j) | v_i,v_j \in V, i \neq j, v_iRv_j \vee v_iCv_j \vee v_iMv_j\}$ 为微博传播网络中边的集合，v_iRv_j 代表用户 v_i 转发过用户 v_j 的微博，v_iCv_j 代表用户 v_i 评论过用户 v_j 的微博，v_iMv_j 代表用户 v_i 在微博中提及过用户 v_j；$B=\{B_{i,j} | (v_i,v_j) \in E\}$ 为微博传播网络中边权集合，$B_{i,j}$ 代表用户 v_i 对用户 v_j 的转发、评论、提及次数的带权之和。

三种行为对微博的传播有着不同的贡献比例，所以 $B_{i,j}$ 应该综合考虑每种行为的贡献比例。本文用 α、β、γ 分别表示转发、评论、提及三种行为的贡献比例，其公式如式（2）所示。

$$B_{i,j} = \alpha R_{i,j} + \beta C_{i,j} + \gamma M_{i,j} \tag{2}$$

其中：$R_{i,j}$、$C_{i,j}$、$M_{i,j}$ 为用户 v_i 对用户 v_j 的转发、评论、提及次数。对于 α、β、γ 的取值本文根据序关系法确定，首先将各个变量的重要程度做成对比较，然后将比较的结果按一定的方式聚合起来，最终经过计算得到 α、β、γ 的值。对 α、β、γ 构建一个判断矩阵 $\boldsymbol{A}$，如式（3）所示。

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} \alpha/\alpha & \alpha/\beta & \alpha/\gamma \\ \beta/\alpha & \beta/\beta & \beta/\gamma \\ \gamma/\alpha & \gamma/\beta & \gamma/\gamma \end{bmatrix} \tag{3}$$

矩阵 $\boldsymbol{A}$ 中的元素 a_{ij} 代表第 i 个变量对第 j 个变量的相对重要性，在矩阵 $\boldsymbol{A}$ 中元素的性质有 $a_{ij}=1/a_{ji}$、$a_{ij}=a_{ik}*a_{kj}$、$a_{ii}=1$。接着根据 Saaty 等人总结出的变量间相对重要性等级表[14]（表 1）结合三种行为的相对重要关系可以对矩阵 $\boldsymbol{A}$ 中的元素赋值。其中，转发行为与评论行为相比处于略微重要到相当重要之间，故取 $a_{12}=4$。转发行为同提及行为相比，处于明显重要与绝对重要之间，故取 $a_{13}=8$。评论关系和提及关系相比，处于同等重要与略微重要之间，故取 $a_{23}=2$。将 a_{12}、a_{13}、a_{23} 代入到式(3)中，得到的最终矩阵如式(4)所示。

$$\boldsymbol{A} = \begin{bmatrix} 1 & 4 & 8 \\ 0.25 & 1 & 2 \\ 0.125 & 0.5 & 1 \end{bmatrix} \tag{4}$$

联立方程组，最后 α、β、γ 的值如式（5）所示。

$$\begin{cases} \alpha = 0.727 \\ \beta = 0.182 \\ \gamma = 0.091 \end{cases} \tag{5}$$

表 1　变量之间的相对重要性

Table 1　Relative Importance between variables

相对重要程度	定义	说明
1	同等重要	两个变量同样重要
3	略微重要	一个变量比另一个变量稍微重要
5	相当重要	一个变量比另一个变量更为重要
7	明显重要	一个变量比另一个变量更为重要，且已有实践证明
9	绝对重要	重要程度可以肯定
2，4，6，8	两个相邻判断的中间值需要折中时采用	

3.2　初始影响力计算

本文从用户的粉丝数、发博数、认证情况三个基本属性入手计算用户的初始影响力值。由于不同用户之间的粉丝数、发博数差异巨大，本文采用一种对数归一化的方法对用户的粉丝数、微博数进行处理，降低数据的跨度。计算公式如式（6）所示。

$$InitInfl(v_i) = \frac{lg(NF(v_i))}{lg(NF_{max})} + \frac{lg(NW(v_i))}{lg(NW_{max})} + Verify(v_i) \tag{6}$$

其中：$NF(v_i)$ 代表真实粉丝数，即用户 v_i 在微博传播网络中的入度；NF_{max} 代表真实粉丝数最多的用户的真实粉丝数；$NW(v_i)$ 代表用户 v_i 的发博数；NW_{max} 代表发博数最多的用户的发博数；$Verify(v_i)$ 代表用户 v_i 的认证情况，Lappas T 等人[15]指出，当用户 v_i 得到了微博认证时 $Verify(v_i)=0.5$，否则 $Verify(v_i)=0$ 的取值最为合适。

3.3　MDIR 算法

MDIR 算法通过合理的策略重新计算了目标用户的影响力，其计算公式如式（7）所示。

$$MDIR(v_i) = (1-d) + d \sum_{v_j \in in(v_i)} ratio(v_j, v_i) * MDIR(v_j) \tag{7}$$

其中：$in(v_i)$ 代表用户 v_i 在微博传播网络中的入度集合；d 是阻尼系数，取值一般为 0.85；$ratio(v_j,v_i)$ 代表用户 v_j 对用户 v_i 的影响力贡献比例，其计算公式如式（8）所示。

$$ratio(v_j, v_i) = \frac{W(v_j, v_i)}{\sum_{v_k \in out(v_j)} W(v_j, v_k)} \tag{8}$$

其中：$out(v_j)$ 为用户 v_j 在微博传播网络中的出度集合；$W(v_j,v_k)$ 代表用户 v_j 对用户 v_k 的微博的传播意愿，由 v_j 对 v_k 的交互频率 BF(v_j,v_k)与兴趣相似度 SIM(v_j,v_k)的乘积表示。

对于用户间的交互频率，可以由微博传播网络 G 的边权集合 B 计算得出，如式（9）所示。

$$BF(v_i, v_j) = \frac{B_{i,j}}{\sum_{v_k \in out(v_i)} B_{i,k}} \tag{9}$$

对于用户间兴趣相似度，本文采用 LDA 模型抽取用户的主题分布向量，而后再基于微博传播网络计算相邻用户间的相似度。首先将用户发布、转发、评论过的历史微博信息聚合成一篇文档，再将“文档—用户”集合作为模型的输入，最后利用 LDA 模型输出每一个用户所对应的主题分布向量。所有用户的主题分布向量集合记做矩阵 $\boldsymbol{DT}$，其中 D、T 分别对应用户数和主题数，矩阵元素 $DT_{i,j}$ 代表用户 v_i 在话题 t_j 上的概率。

基于“用户—主题”矩阵 $\boldsymbol{DT}$，本文采用 KL 距离计算用户间相似性，KL 距离是描述两个概率分布 P、Q 差异的一种方法，其计算公式如式（10）所示。

$$KL(P||Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)} \tag{10}$$

根据式（10）可以看出，两概率分布之间越相似则它们

的 KL 离散度越小，且 KL 离散度不具有对称性，即 $KL(P\|Q) \neq KL(Q\|P)$。由此，为了更适当地表示微博用户间的相似度，本文先对 KL 距离取平均而后再取倒数，如式（11）所示。

$$SIM(v_j,v_i)=\frac{2}{KL(DT_i\|DT_j)+KL(DT_j\|DT_i)} \tag{11}$$

其中：DT_i 与 DT_j 分别为矩阵 ***DT*** 的第 i 行与第 j 行，表示用户 v_i 与用户 v_j 的话题分布向量。

以下是 MDIR 算法的主要处理过程：

输入：微博传播网网络 $G=(V,E,B)$，阻尼系数 d，阈值 ε。

输出：所有用户的影响力集合 S。

```
//微博传播网络边权值初始化
for B_{i,j} ∈ B
    B_{i,j} = αR_{i,j} + βC_{i,j} + γM_{i,j}
end for
//用户 v_i 的影响力值初始化
for v_i ∈ V
    MDIR(v_i) = lg(NF(v_i))/lg(NF_max) + lg(NW(v_i))/lg(NW_max) + Verify(v_i)
end for
S.Init() //初始化集合 S
//用户影响力计算，当所有用户的影响力值都收敛时，迭代结束
while |S| < |V|
while (v_j,v_i) ∈ E
    P_0 = MDIR(v_i) //用户 v_i 上一轮计算的影响力值
    BF(v_j,v_i) = B_{i,j} / Σ_{v_k∈out(v_j)} B_{i,k}
    SIM(v_j,v_i) = 2 / (KL(DT_i||DT_j) + KL(DT_j||DT_i))
    ratio(v_j,v_i) = W(v_j,v_i) / Σ_{v_k∈out(v_j)} W(v_j,v_k)
    MDIR(v_i) = (1-d) + d Σ_{v_j∈in(v_i)} ratio(v_j,v_i) * MDIR(v_j)
    P = MDIR(v_i)  //用户 v_i 当前轮次计算的影响力值
    if | P-P_0 | ≤ ε
        S.add(v_i,P)      //保存用户 v_i 的影响力
    end if
end while
end while
return S
```

根据式（8）可以构造概率转移矩阵 ***M***，从而 MDIR 算法的求解过程可以转换成一个 Markov 过程。Markov 过程的收敛条件有：a) 矩阵 ***M*** 为随机矩阵；b) 矩阵 ***A*** 为不可约矩阵；c) 矩阵 ***M*** 为非周期矩阵。根据式（8）的表述，矩阵 ***M*** 可直接满足收敛条件 a)和 b)；对于收敛条件 c)，本文在实验开始前会对用户进行筛选，使得构建的微博传播网络是一个强连通图，可以保证矩阵 ***M*** 的不可约性。综上所述，本文所提出的 MDIR 算法是可收敛的。

4　实验结果与分析

4.1　数据集与实验环境

4.1.1 数据集选取

本文以新浪微博为作为数据源，爬取了 2014 年 5 月份新浪微博一个月内 12 个热门话题下的部分用户微博信息作为研究数据。由于爬取到的用户信息繁杂、冗余信息过多，本文过滤掉发博数少于 10、关注用户数少于 10 的用户，由筛选后的数据所构建的微博传播网络的相关数据如表 2 所示。

表 2　微博传播网络相关数据

Table 2　Related data of microblog propagation network

信息	数据
总节点数	81346
总边数	2712345
平均出度数	18.37
平均入度数	17.12
总微博数	1091461

4.1.2 实验环境

由于在构建微博传播网络与 MDIR 算法迭代计算时所耗费的时间开销较大，本文对这两个过程进行了基于 MapReduce 的并行化设计。实验使用四台 PC 机搭建 Hadoop 集群，每台机器的操作系统均为 64 位 CentOS-7。具体集群概况如表 3 所示。

表 3　实验集群概况

Table 3　Overview of experimental clusters

IP 地址	主机名	集群角色
172.168.21.5	Master	NameNode SecondNameNode ResourceManager
172.168.21.6	Slave01	DataNode
172.168.21.7	Slave02	DataNode
172.168.21.8	Slave03	NodeManager

4.2　对比实验与评价标准

4.2.1 对比实验

为了让实验结果更有说服力，本文选取了目前较为流行或经典的用户影响力度量算法作为对比实验。首先选取的是 PageRank 算法，由于 MDIR 算法是基于 PageRank 的改进算法，用原始的 PageRank 算法做对比更会凸显该算法的优点；其次选取的是 BWPR 算法，齐超等人[8]基于 PageRank 算法提出了一种 BWPR 算法计算用户影响力，其主要改进是基于用户间交互行为来确定粉丝用户影响力的分配因子；然后是 TwitterRank 算法，Weng 等人[9]提出了针对 Twitter 平台的 TwitterRank 算法，该算法主要是在 PageRank 算法的基础上融合了用户的兴趣相似度，虽然该算法是基于 Twitter 平台的，但是 Twitter 平台与新浪微博平台在结构上是一脉相承的，所以将 TwitterRank 算法推广在微博平台也具有一定意义；最后两种分别是基于用户粉丝数与发博数的排名算法。

4.2.2 评价标准

在实际环境中微博用户影响力的衡量标准众多，难以给定一个统一标准。本文采用丁兆云等人[16]提出的 M 折交叉验证的方法，分别验证了算法的准确率、召回率和 F 值。首先求出实验中的五种排序算法与本文提出的 MDIR 算法各自所计算出的 Top-K 影响力用户集合 I_x，然后构造标准排序集合 I_M 为任意 M (1<M≤6)种算法都投票为正确的结果。集合 I_M 的算术描述如式（12）所示。

$$I_M=\bigcup_{x\in Combine(6,M)}\left(\bigcap_{i=1}^{M} I_{x_i}\right) \tag{12}$$

其中：*Combine(6,M)*是从六种算法中选取 *M* 种算法的组合数。例如，给定四种算法 A、B、C、D 得到的 Top-K 个影响力用户集合为 I_A、I_B、I_C、I_D，假设 *M*=2，则标准集合 I_M 的构成如式（13）所示。

$$I_M=(I_A\cap I_B)\cup(I_A\cap I_C)\cup(I_A\cap I_D)\cup(I_B\cap I_C)\cup(I_B\cap I_D)\cup(I_C\cap I_D) \tag{13}$$

其中：算法 A 的准确率 P_A 的计算公式如式（14）所示。

$$P_A = \frac{|I_A \cap I_M|}{|I_A|} \tag{14}$$

算法 A 的召回率 R_A 的计算公式如式（15）所示。

$$R_A = \frac{|I_A \cap I_M|}{|I_M|} \tag{15}$$

算法 A 的 F_A 值的计算公式如式（16）所示。

$$F_A = 2 * \frac{P_A * R_A}{P_A + R_A} \tag{16}$$

4.3　算法有效性验证

同时对六种算法在 M={2,3,4}的情况下进行交叉验证。由于 $M \geq 5$ 时，标准集合内的元素较少，各算法的准确率与召回率比较相似，故予以忽略。针对 M=2、3、4 的三种情况对六种算法所得的 Top-K (K 分别值取 100、200、...、1000) 影响力用户的准确率、召回率及 F 值进行比较，实验结果表明 MDIR 算法在三种衡量指标下都有不错效果。

4.3.1 准确率验证

在用户规模为 Top-K 的情况下，准确率表示目标算法正确计算出的 Top-K 用户的个数与用户数 K 的比值。由图 1 所示的三组实验结果表明，MDIR 算法在不同的用户规模 K 与交叉折数 M 下的准确率都优于其他对比算法。其中当 M=3、4 时，集合 I_M 中的用户较少，从而使得任意算法的结果集与标准集的相交的元素也较少，故整体的准确率相较于 M=2 的时候要低一些。

4.3.2 召回率验证

在规模为 Top-K 影响力用户中，召回率为目标算法“正确”计算出的 Top-K 用户的个数与标准集中影响力用户个数的比值，反映了微博中影响力用户被发现的程度。六种算法分别在 M=2、3、4 的情况下 Top-K 影响力用户的召回率分布如图 2 所示。实验表明，MDIR 算法在 M 的三种取值下均有不错的表现，在 M=3 的情况下 MDIR 算法区分度尤为明显。因为召回率由 $|I_A \cap I_M|$ 和 $|I_M|$ 共同决定，所以当 M 增加时两者都随之增加，召回率变化不明显。

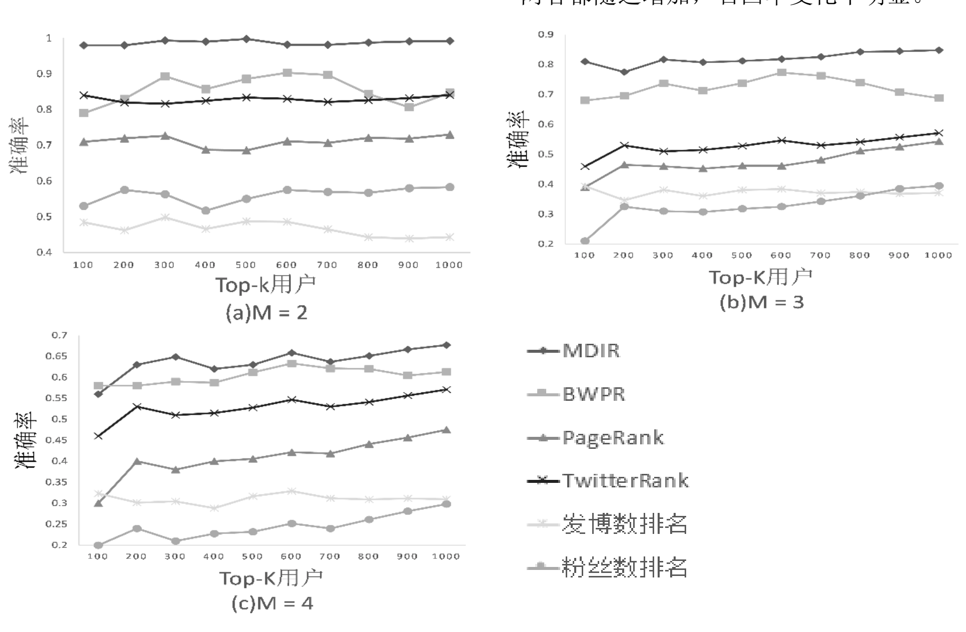


图 1　各算法在交叉验证中的准确率

Fig. 1　Precision of algorithms in cross-validation

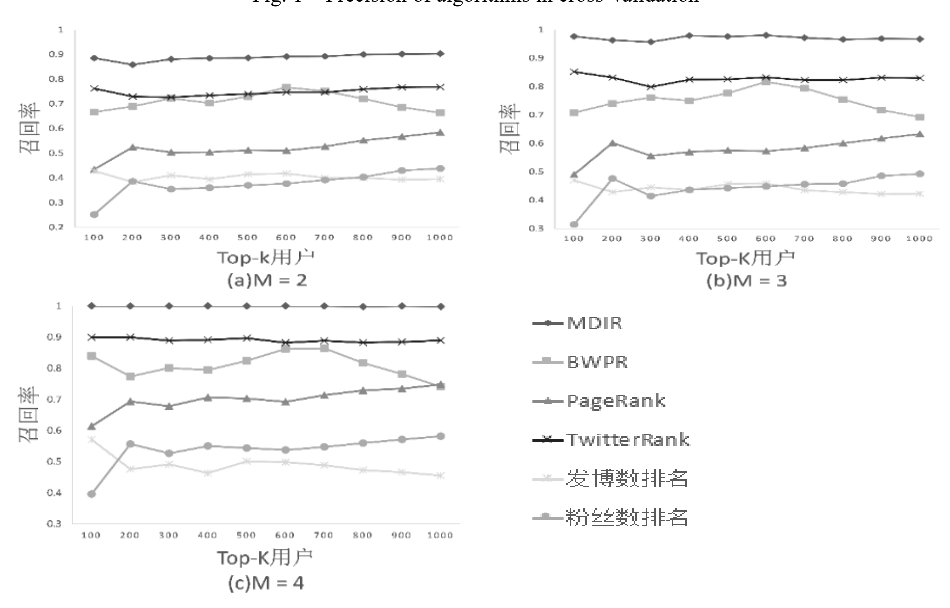


图 2　各算法在交叉验证中的召回率

Fig. 2　Recall of algorithms in cross-validation

4.3.3 F 值验证

F 值综合考虑召回率与准确率，反映算法整体召回率与准确率的程度。各算法在交叉验证中的 F 值如图 3 所示。由图 3 可知，本文提出的 MDIR 算法在三组实验中有着明显的优势；同时可以看出只基于用户交互行为的 BWPR 算法在统计规模增加的情况下效率有所下降，这是因为用户的影响力还与用户的基本属性与博文内容有关；基于用户博文相似度的 TwitterRank 算法效率虽然会随着统计规模的增加出现上升的趋势，但是其没有结合用户的交互行为，所以效率一直不佳；PageRank 算法和用户粉丝数排名的变化趋势很相近，足以说明 PageRank 很大程度依赖于用户的粉丝数，这也是原始 PageRank 算法的局限所在；用户发博数的排名与用户粉丝数排名由于考虑因素的单一，致使整体性能很差；而本文提出的 MDIR 算法则综合考虑了用户自身基本属性、用户间的交互行为以及博文内容三个维度的因素，并且有针对性地将其融入原始的 PageRank 算法，使得其最终计算的结果在准确率、召回率、F 值上均优于其他对比算法。

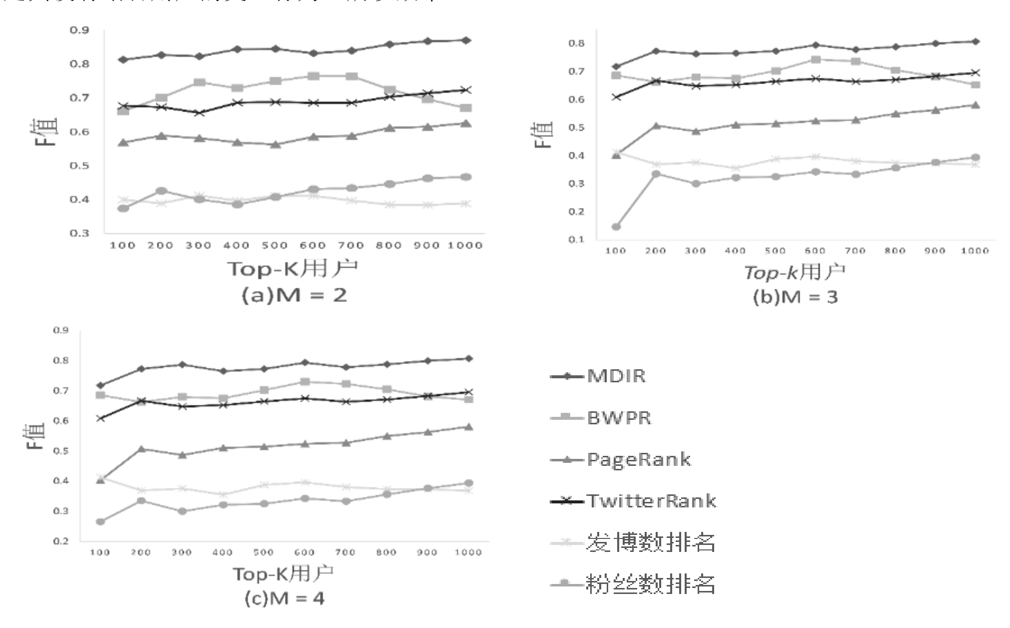


图 3　各算法在交叉验证中的 F 值

Fig. 3　F-Measure of algorithms in cross-validation

4.3.4 收敛性验证

SF-UIR 算法是王顼[6]等人于 2018 年提出的微博用户影响力度量算法，该算法在 PageRank 的基础上加入了用户自身行为的特征与网络拓扑结构中的粉丝用户特征，解决了传统 PageRank 算法客观性差、影响力传递比例分配均匀的缺点，并且通过实验验证了该算法计算出的影响力排序结果的全面性与真实性。为了进一步说明本文提出的 MDIR 算法的适用性，本文将对比两个算法收敛速度。

本文仍用表 2 中介绍的数据集，分别对两种算法进行并行化，在相同收敛阈值的情况下计算用户影响力值，并且跟踪记录计算过程中的迭代次数。结果发现，MDIR 算法使影响力值收敛的迭代次数为 58 次，用时 34 min。对于 SF-UIR 算法而言，其迭代次数为 65 次，用时 39 min。实验结果说明了本文提出的 MDIR 算法有较好的收敛性。

4.4　算法时间效率验证

根据 3.3 节中 MDIR 算法的相关介绍可以看出，由于要计算用户间的影响力贡献比例 $ratio(v_j,v_i)$，所以本文提出的 MDIR 算法相较于 PageRank 算法的时间复杂度有所增加。但 $ratio(v_j,v_i)$ 的计算简单，仅涉及微博文本内容与用户交互关系，这些因素都可以在实验之前的预处理步骤中提取得到。另外，MDIR 算法可以达到高影响力用户的影响力值更快累计，低影响力用户的影响力值更快趋于收敛的效果。为了验证并行化之后的 MDIR 算法的时间效率，本文比较了单机串行的 MDIR 算法与基于 MapReduce 的并行化 MDIR 算法在处理相同规模数据时，从读入数据到最后收敛所消耗的时间。根据不同的用户规模设计了八次对比测试，其结果如图 4 所示。从图中不难看出，基于并行化的 MDIR 算法在处理规模较小的 1 万条用户数据时，所花时间为 10 min，略低于串行 MDIR 算法所消耗的 7 min。然而从 3 万条用户数据开始，并行化 MDIR 算法的执行时间的变化较为平稳，但是串行 MDIR 算法几乎呈指数型增长。当达到的 8 万用户的数据规模时，并行化算法所需的时间几乎为串行算法的一半。如果数据量持续增大，那么串行算法会因为消耗的内存过多，引起程序的异常退出，而经过并行化后的 MDIR 算法依然会保持不错的性能。

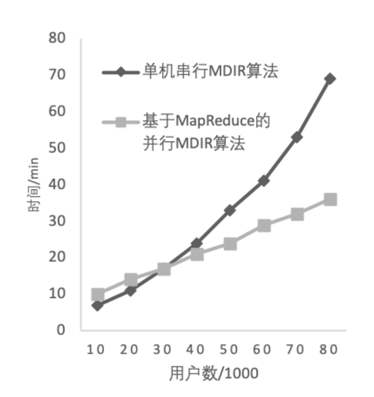


图 4　串、并行 MDIR 算法时间效率对比

Fig. 4　Time efficiency comparison of algorithm between serial MDIR and parallel MDIR

5　结束语

本文通过分析传统 PageRank 算法度量用户影响力存在的问题，从三个维度进行了有针对性的特征选取，进而提出了 MDIR 算法。相比当前的相关研究，MDIR 算法融入了更

多的影响因素，其考虑用户自身的基本属性使计算结果更为客观，考虑用户间的交互行为与博文内容使计算结果更为合理，考虑网络拓扑结构规避了不必要的计算并且获得更有效的结果。通过实验也可以看出，本文提出的 MDIR 算法在多种影响力评价指标上取得了良好的效果。

在下一步的研究中，将会首先考虑用户在不同话题中的影响力，挖掘在多个话题中都具有高影响力的意见领袖。其次，微博中大量的“水军”会对计算结果产生一定的干扰，需要在度量算法中加入对微博“水军”的识别技术，使得最终的计算结果更客观。

参考文献：

[1] CNNIC. 第 41 次《中国互联网络发展状况统计报告》[R]. 北京：中国互联网络信息中心. 2018. (CNNIC. The 41st Statistical report on the development of China's Internet network [R]. Beijing: China Internet Network Information Center, 2018.)

[2] 王晨旭，管晓宏，秦涛，等. 微博消息传播中意见领袖影响力建模研究 [J]. 软件学报，2015，26 (6)：1473-1485. (Wang Chenxu, Guan Xiaohong, Qin Tao, *et al.* Modeling on opinion leader’s influence in microblog message propagation and its application [J]. Journal of Software, 2015, 26 (6): 1473-1485.)

[3] Cha M, Haddadi H, Benevenuto F, *et al.* Measuring user influence in Twitter: the million follower fallacy [C]// Proc of International Conference on Weblogs and Social Media. Carlifornia: AAAI Press, 2010: 14.

[4] Mao Guojun, Zhang Jie. A PageRank-based mining algorithm for user influences on micro-blogs [C]// Proc of Pacific Asia Conference on Information Systems . Berlin: Springer, 2016.

[5] 张昊，刘功申，苏波. 一种微博用户影响力的计算方法 [J]. 计算机应用与软件, 2015, 32 (3): 41-44. (Zhang Hao, Liu Gongshen, Subo. A computer method for microblog ging users influence [J]. Computer Applications and Software, 2015, 32 (3): 41-44.)

[6] 王顶，徐军，段存玉，等. 基于 PageRank 的用户影响力评价改进算法 [J]. 哈尔滨工业大学学报, 2018, 50 (5): 60-67. (Wang Ding, Xu Jun, Duan Cunyu, *et al.* Improved user influence ecaluation algorithm based on PageRank [J]. Journal of Harbin insititude of Technology: Social Sciences Edition, 2018, 50 (5): 60-67.)

[7] 孙红，左腾. 基于 PageRank 的微博用户影响力算法研究 [J]. 计算机应用研究, 2018, 35 (4): 1028-1032. (Sun Hong, Zuo Teng. Research on algorithm of micro-blog user influence based on PageRank [J]. Application Research of Computers, 2018, 35 (4): 1028-1032.)

[8] 齐超，陈鸿昶，于洪涛. 基于用户行为综合分析的微博用户影响力评价方法 [J]. 计算机应用研究, 2014, 31 (7): 2004-2007. (Qi Chao, Chen Hongxu, Yu Hongtao. Method of evaluating micro-blog users’ influence based on comprehensive analysis of user behavior [J]. Application Research of Computers, 2014, 31 (7): 2004-2007.)

[9] Weng Jianshu, Lim E P, Jiang Jing, *et al.* TwitterRank: finding topic-sensitive influential twitterers [C]// Proc of ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 261-27.

[10] 师亚凯，马慧芳，张迪，等. 融合用户行为和内容的微博用户影响力方法 [J]. 计算机应用研究, 2016, 33 (10): 2906-2909. (Shi Yakai, Ma Huifang, Zhang D, *et al.* Microblog user influence algorithm based on user behavior and content [J]. Application Research of Computers, 2016, 33 (10): 2906-2909.)

[11] Brin S, Page L. Repeint of: The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks, 2012, 56 (18): 3825-3833.

[12] Wang Chi, Chen Wei, Wang Yajun, *et al.* Scalable influence maximization for independent cascade model in large-scale social networks [J]. Data Mining and Knowledge Discovery, 2012, 25 (3): 545-576.

[13] 李志宏，庄云蓓. 基于 PageRank 算法的双维度微博用户影响力实时度量模型 [J]. 系统工程，2016，34 (2)：128-137. (Li Zhihong, Zhuang.Yunbei. Rela-time measurement model of the influence of micro-blog users with double dimensions based on PageRank algorithm [J]. System Engineering, 2016, 34 (2): 128-137.)

[14] Saaty T. Fundamentals of the analytic network process-multiple networks with benefits [J]. Journal of System Science and System Engineering, 2004, 13 (3): 348-379.

[15] Lappas T, Terzi E, Gunopulos D, *et al.* Finding effectors in social networks [C]// Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 1059-1068.

[16] 丁兆云，周斌，贾焰，等. 微博中基于多关系网络的话题层次影响力分析 [J]. 计算机研究与发展，2013，50 (10)：2155-2175. (Ding Zhaoyun, Zhou Bin, Jia Yan, *et al.* Topical influence analysis based on the multi-relational network in microblogs [J]. Journal of Computer Research and Development, 2013, 50 (10): 2155-2175.)